Finding Significant Predictors of Systolic Blood Pressure

Evan Turkon, Ireri Avila, Yazhu Jiang

October 2023

Introduction

In the context of the rising prevalence of cardiovascular diseases, this report seeks to explore the factors influencing systolic blood pressure values. Our study revolves around the application of linear regression to understand how various elements contribute to blood pressure levels. Specifically, we aim to establish a relationship between blood pressure and a range of independent variables, including but not limited to age, BMI, race, and gender. By leveraging regression analysis, we aim to unearth patterns, quantify impacts, and provide valuable insights.

The Dataset

All of the data being explored in this project came from The National Health and Nutrition Examination Survey (NHANES) [1] which is a program by the National Center for Health Statistics (NCHS), which is a partner with the Centers for Disease Control and Prevention (CDC). NHANES is designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines both interviews and physical examinations.

While many possible predictors were tested in this analysis, table 1 describes the features included in the final model.

Variable	Description	Data Type
blood pressure	(Dependant Variable) Systolic Blood Pressure	Numeric (Float)
race	(Black, Mexican American, Multi Racial, Other Hispanic, White)	Categorical
gender	(Male, Female)	Categorical
age_height	Engineered feature: log(age*height)	Numeric (Float)
cocaine_uses	number of times participant used cocaine	Numeric (Integer)
bmi	Body Mass Index	Numeric (Float)

Table 1: Variables Included in Final Linear Regression Model

Exploratory Analysis

The obtained data contains 2642 records with 32 columns. The columns correspond to the questions from the survey.

We cleaned the data by removing rows where *blood pressure*, and *height* had missing values. The rest of the variables where analyzed in three groups. Those related to:

- Physical measures
- Nutritional Intake
- Substance Use

When checking for the individual relation between nutritional variables and blood pressure, the results were almost identical among all of them. Further we analyzed the nutritional intake as an aggregate for *water, carbohydrates, fiber, protein*, and *sugar*. This showed that all of them cancelled each other out, thus we decided not to consider them among the variables for our model.

We then checked for those variables related to substance use. These plots are shown in fig. 1. Most of the relations appeared similar with a cluster of points to the left, indicating little to no change when an increase in the x_axis occurred with respect to the response variable. *Cocaine use* may be an interesting variable to explore as there is some difference between the observed ranges for blood pressure as a person increases their cocaine use.

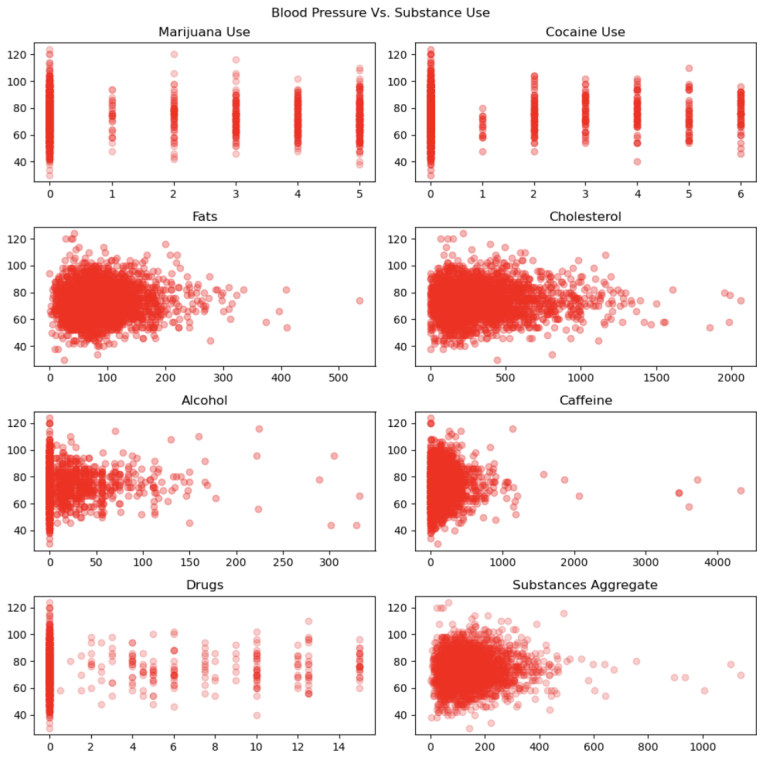

Figure 1: Plots for relationship between substance use related predictors, and response variable.

Finally we explored the different relationships among our possible predictors for physical measures with the response variable, including the previously plotted *cocaine use* as a comparison among them. This is shown in Fig. 2.

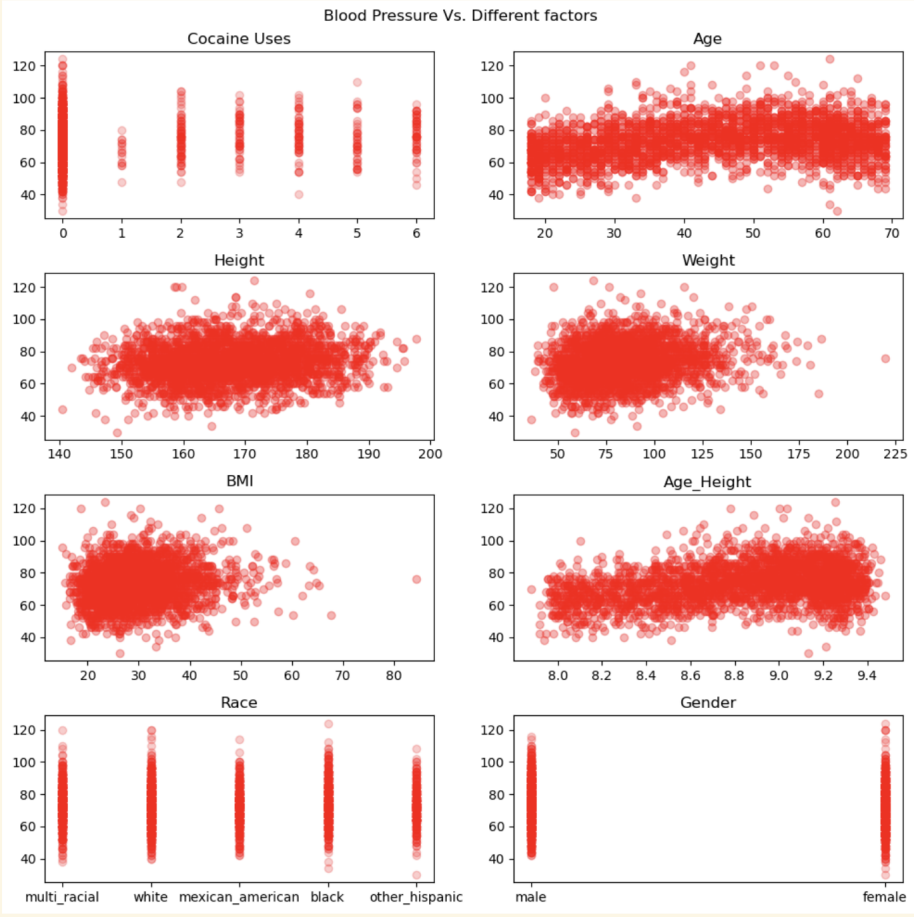

Figure 2: Plots for relationship between physical measures predictors, and response variable.

From this we can observe that there is no clear pattern among these relations. Furthermore there appears that the relationships do not appear to approach a single value but, there appears to be a range of values in *Blood pressure* that correspond to single values in the corresponding predictor. This is more clearly seen in Fig.2 for our categorical predictors: *Cocaine Uses, Race* and *Gender*.

It is further seen that individual variables, like *age, height*, and *weight* show almost a constant pattern for blood pressure while we observe more insight on a possible linear pattern for the composite variable *age_height*. This is also compared with our second composite variable *bmi*.

Bmi takes the *weight* and *height* of the person for its calculation,

Model Selection

Table 2 compares model summaries between a preliminary hypothesized model **Model 1**, and a fine tuned model **Model 2**. The dependent variable blood pressure stays constant but in model 2 the logarithm is taken which helped to linearize the relationship between it and the predictors, this is shown in the higher adjusted R^2 of Model 2. A Breusch Pagan test confirms this and is discussed further in the model diagnostics section. While 2 of the races were not statistically significant, race was included in this model due to the significance of the other races and the outcome of the F-test in the anova type=1 table where the null hypothesis is rejected and it can be concluded that race is a significant predictor of blood pressure given all other predictors in the model. It was found that age and height had a better linear relationship with blood pressure when the logarithm of their product is included in the model instead of including them separately as seen in the difference in Model 1.

Variable	Model 1	Model 2
dependent variable	blood pressure	log(blood pressure)
intercept	9.7377 (***)	3.1146 (***)
race(mexican_american)	-0.4769 (ns)	-0.0089 (ns)
race(multi_racial)	1.2544 (*)	0.0160 (*)
race(other_hispanic)	-0.9659 (ns)	-0.0177 (*)
race(white)	-0.7945 (ns)	-0.0119 (ns)
gender(male)	1.5693 (***)	0.0394 (***)
age	8.8693 (***)	
height	0.1483 (***)	
bmi	0.1701 (***)	0.0024 (***)
alchohol	0.0056 (ns)	
marijuana_use	0.0503 (ns)	
min_vig_rec	-0.0023 (ns)	
age_height		0.1230 (***)
cocaine_uses		0.0078 (***)
$adj - R^2$	0.112	0.17

Table 2: Model Selection (Significance Levels: (***) ≤ 0.01, (**) ≤ 0.05, (*)≤ 0.1, (ns) = not significant)

Variable	F-test p-value
gender	1.617591e-23
race	1.382305e-02
age_height	3.422415e-75
bmi	4.259582e-10
cocaine_uses	1.924606e-04

Table 3: ANOVA(typ=1) Table for Model 2

Regression Analysis

While all of the predictors in Model 2 do pass the F-test for significance in predicting systolic blood pressure, many of the coefficients of these predictors are very small, even accounting for the logarithmic transformation. As described in the Exploratory Analysis section systolic blood pressure doesn't fluctuate much so these small coefficients can be explained and while they may be small they are extremely significant. This low volatility of blood pressure also helps to explain the low $adj - R^2$. The coefficient of the engineered feature 'age_height' is difficult to interpret but does reveal a strong positive relationship between the logarithm of the product of age and height. The signs of the other coefficients for the numeric variables are as expected for the model, blood pressure has a positive relationship with bmi, the number of times you've used cocaine.

Model Diagnosis

We checked for the assumption of constant variance by making a plot of the fitted values against the residuals, and got the following:

Figure 3: Non Log Transformation VS. Log Transformation

In our analysis of the linear regression model (Model 1), which did not involve the use of log transformation on the variables, we conducted a Breusch-Pagan test to assess the presence of Heteroscedasticity. The test resulted in a p-value close to zero, indicating evidence of heteroscedasticity and, consequently, non-constant variance in the residuals.

To address this issue, we applied a log transformation to both the independent and dependent variables. The log transformation is a common technique for stabilizing variance and addressing heteroscedasticity. After implementing this transformation, we re-evaluated the presence of heteroscedasticity using the Breusch-Pagan test. Remarkably, the p-value increased to approximately 0.67, suggesting that heteroscedasticity no longer exists in the transformed model.

This transformation not only mitigated the issue of heteroscedasticity but also improved the model's ability to meet the assumptions of linear regression. As a result, the log-transformed model provides a more suitable framework for making reliable inferences and predictions.

Overall, our analysis underscores the importance of addressing heteroscedasticity in regression modeling and demonstrates how log transformation can be a valuable tool in achieving more robust and accurate results.

We then analyzed the VIF. This analysis reveals distinct patterns in variable correlations within the model.

	VIF Factor	features
0	1573.429336	Intercept
1	2.936607	salt_type[T.lite]
2	33.850093	salt_type[T.no_extra_salt]
3	35.598283	salt_type[T.ordinary]
4	2.773899	salt_type[T.salt_substitute]
5	1.151724	gender[T.male]
6	1.595490	race[T.mexican_american]
7	1.784395	race[T.multi_racial]
8	1.372923	race[T.other_hispanic]
9	1.820172	race[T.white]
10	1.003345	SEQN
11	1.125264	bmi
12	3.738428	protien
13	8.972718	carbs
14	5.350533	sugar
15	2.103572	fiber
16	3.242813	fats
17	2.205944	cholesterol
18	4.012873	sodium
19	1.109057	caffeine
20	1.035506	water
21	1.256554	age
22	1.432424	household_person_count
23	1.349794	children_u5_household_count
24	1.966262	doc_diabetes
25	2.032197	risk_diabetes

Figure 4: VIF Coefficients for predictors in model

Specifically, *salt_type[T.no_extra_salt]* and *salt_type[T.ordinary]* exhibit strong correlations with other

predictors.

Variables indicating race present a moderate correlation, while SEQN appears to be uncorrelated with any other predictors. In the nutritional segment, variables such as *bmi, caffeine, water, age, household_person_count* and *children_u5_household_count* display negligible correlations with other factors.

Notably, *carbs* demonstrate pronounced multicollinearity. Additionally, both *doc_diabetes* and *risk_diabetes* show a moderate correlation with other predictors. However, despite these observations, the model does not seem to suffer from severe multicollinearity issues overall.

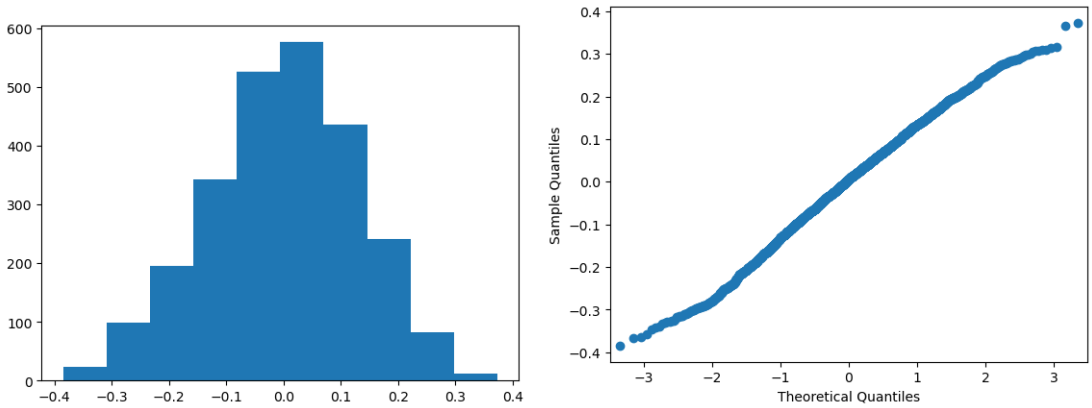

Figure 5: Histogram and QQ-plot of Residuals

Upon visual examination of both the histogram and the QQ plot of the residuals, it becomes evident that they exhibit a strong adherence to a normal distribution. This observation suggests that the residuals fulfill the normality assumption required for linear regression analysis.

In our analysis, we employed two methods, external studentized residuals and Cook's distance, to identify influential points within the dataset, shown in fig 4. Our criteria for identifying these points involved using threshold values derived from the t-distribution. Specifically, we set the threshold for external studentized residuals based on a t-distribution with parameters alpha = 0.05 and degrees of freedom equal to n-p-1, where 'n' represents the sample size and 'p' signifies the number of predictors. For Cook's distance, the threshold was set at 4/n.

By cross-referencing the influential points detected by both methods, we were able to identify their intersection. Prior to the removal of influential points, the coefficient of determination R^2 stood at 0.136. However, following the removal of these points, we observed a substantial increase in R^2, which rose to 0.173.

This enhancement in the R^2 value underscores the significance of addressing influential points in regression analysis. By identifying and removing these influential observations, we were able to obtain a more accurate and reliable linear regression model, thereby improving its predictive performance.

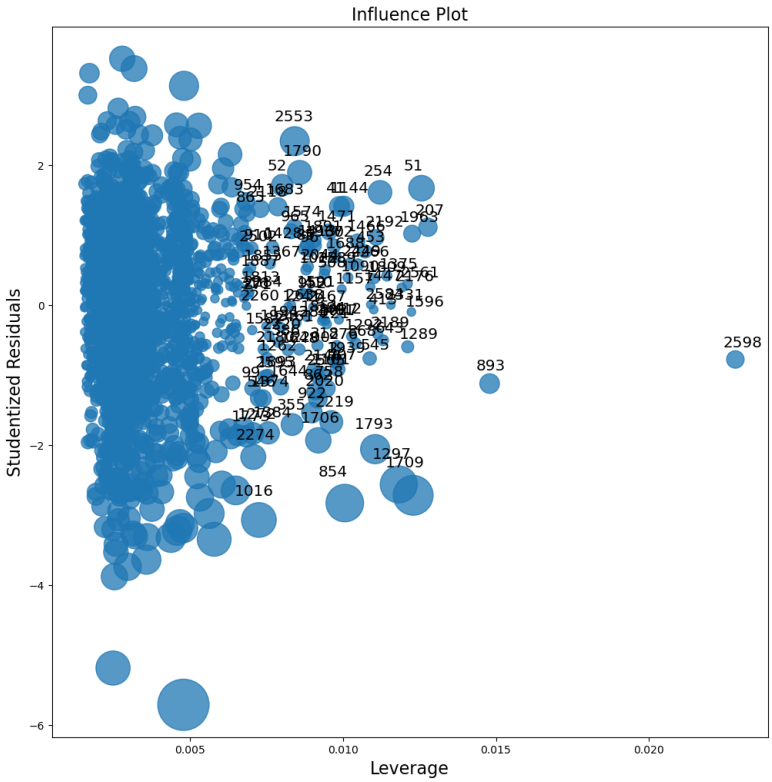

Figure 6: Influence Plot

Results

Our final model with an $AdjR^2 = 0.17$ is:

$$\begin{aligned} Log(Blood\_pressure) = \\ & 3.1146 - 0.0089Race[mex\_am] \\ & + 0.016Race[multi] - 0.0177Race[other\_hisp] \\ & - 0.0119Race[white] + 0.0394Gender[male] + 0.0024Bmi \\ & + 0.1230Age\_height + 0.0078Cocaine\_use \end{aligned} \tag{1}$$

There is no heteroscedasticity in our model so our test results are certain for significance of predictors used. We also observed from our QQ plot that there does appear to be a normal distribution as the line maintains its linear pattern. However, given our low $AdjR^2$ this model is not the best for predictions but helps with inferences on systolic blood pressure.

Further it is observed that blood pressure is a range of accepted values, this can be seen in fig **??** where a line is observed but going directly in the y_axis direction instead of sharing increments with all the values in the x_axis.

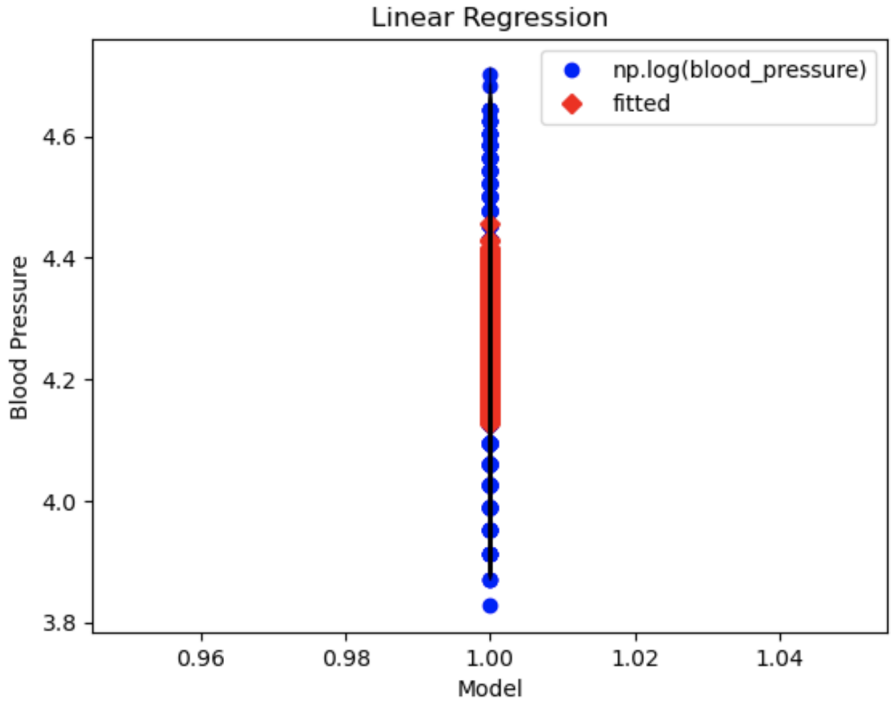

Figure 7: Linear Model

Conclusion

We can conclude that although there is some influence from these factors being considered in our model, fitting a linear model for blood pressure shows that this approach is not the best at explaining the change in blood pressure. Further analysis is needed to either create new composite variables or trying a more complex model.

There are also additional factors out of the scope of the survey that could be better at predicting or understanding the behaviour of blood pressure. Family information, as well as medical history could be a set of parameters to explore.

Blood pressure is a range of values for every individual, and while nutritional factors may not heavily impact on whether you have high or blood pressure, they are significant to it. One example of this is that an increase in *bmi* and *cocaine use* will increase the individual's blood pressure. As well as a higher blood pressure among males than that of females.

References

[1] Center for Disease Control and Prevention. National health and nutrition examination survey. `https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017`, 2017.